Vocabulary Management in Support of Information Services

Rebecca Guenther
Network Development & MARC Standards Office,
Library of Congress
rgue@loc.gov

Semantic-Interoperability (ELS/SenSOC) Summer Workshop July 22, 2010

Outline of presentation

- Types of controlled vocabularies
- Vocabularies maintained at LC
- What is SKOS?
- What is Linked Data?
- id.loc.gov vocabulary services
- How it's being used
- Publishing LC's vocabularies

Why establish controlled vocabularies?

- Control values that occur in metadata
- Reduce ambiguity
- Control synonyms
- Document and publish for reuse
- Test and validate terms
- Establish formal relationships among terms (where appropriate)

Types of Controlled Vocabularies used in metadata standards

- Lists of enumerated values
  - Lists of terms from pull-down menu
  - Enumerated values in XML schema
- Code lists (e.g. language, country)
  - Standardizes the code, not the name
- Taxonomies
- Formal Thesauri
  - Controlled vocabularies with relationships between terms
- Locally controlled enumerated lists

Thesauri

- A *thesaurus* is a controlled vocabulary with multiple types of relationships

Example:

Rice
UF Paddy
BT Cereals
BT Plant products
NT Brown rice
RT Rice straw

Standards maintained at LC that contain controlled vocabularies

- LCSH/NAF
- Thesaurus of Graphic Materials
- MARC Code lists: GACs, countries, languages
- ISO 639-2 and ISO 639-5 (language codes)
- Enumerated lists in XML schemas
  - MODS enumerated values (XML descriptive metadata schema)
  - METS enumerated values (XML information package)
  - MIX (Technical metadata for digital still images)
- PREMIS controlled vocabularies (Preservation metadata)
- Others…

“Linked Data”

- A feature of the “Semantic Web” where links are made between resources
- Goes beyond hypertext links (i.e. between web pages) but between any kind of object or concept
- From Wikipedia: "a term used to describe a method of exposing, sharing, and connecting data via dereferenceable URIs on the Web”
- Users can use links to find similar resources and aggregate results

Simple Knowledge Organization System (SKOS)

- RDF application used to express knowledge organization systems such as thesauri, taxonomies and the concepts within.
- SKOS has a defined element set which is particularly relevant for controlled vocabularies
- Relationships between concepts in a concept scheme can be expressed (e.g. broader, narrower) and between concepts in different schemes
- Having a dereferencable URI for concepts and their concept schemes enhances the ability to provide web services for consumers of these standards

The SKOS data model (Classes)

- ConceptSchemes (e.g., published vocabularies, thesauri, code lists, etc.)
- **Concepts** (individual entries or terms within the larger vocabulary)
- Collections (logical groupings of Concepts)

SKOS concepts

- Labeling properties: **prefLabel, altLabel, hiddenLabel, notation**
- Annotation properties: **note, historyNote, scopeNote, changeNote, editorialNote, example, definition**
- Associative properties: **broader, narrower, related,_broadMatch, narrowMatch, closeMatch, exactMatch, minorMatch, majorMatch** (match properties go to Concepts in external ConceptSchemes)

Reasons for developing a web service for vocabularies

- Facilitate development and maintenance process for vocabularies
- Make controlled lists openly available
- Provide comprehensive information about controlled terms
- Experiment with semantic web technologies and linked data
- Expose vocabularies to wider communities

Introducing id.loc.gov

- Library of Congress Authorities & Vocabularies service: http://id.loc.gov
- Allows both human-oriented and programmatic access to LC-promulgated authorities and vocabularies.
- First offering was Library of Congress Subject Headings in April 2009
- Additional vocabularies added April 2010
- Data is continuously updated

Available vocabularies

- LCSH
- Thesaurus for Graphic Materials
- MARC Code List for Relators
- Preservation events
- Cryptographic hash functions (METS, PREMIS and MIX)
- Preservation level role

URIs in id.loc.gov

- Interaction with any given individual term and vocabulary is with its URI
- Some examples of URIs:

http://id.loc.gov/vocabulary/relators/art

http://id.loc.gov/vocabulary/graphicMaterials/tgm005222

http://id.loc.gov/vocabulary/preservationEvents/migration

http://id.loc.gov/authorities/sh85063136

- Known-label searches: use when you know the label but not the identifier

http://id.loc.gov/vocabulary/relators/label/artist

http://id.loc.gov/authorities/label/hunting%20dogs

Other features of id.loc.gov

- Can search terms in one or multiple vocabularies:

Hunting dogs in LCSH and TGM

- Visualizations
- Links to similar concepts in other vocabularies (e.g. Rameau)

ID: how it's being used
University of Pennsylvania

Medicine, Botanic

Here are entered works on a 19th century system of medicine developed by Samuel Thomson and based on the use of plant remedies.

Broader terms:

- Alternative medicine
- Medicine

Used for:

- Botanic medicine
- Thomsonianism

Filed under: Medicine, Botanic

A Guide to Health, Being an Exposition of the Principles of the Thomsonian System of Practice, and Their Mode of Application in the Cure of Every Form of Disease (1846 edition), by Benjamin Colby (text and PDF with commentary at swsbm.com)

Life and Medical Discoveries of Samuel Thomson, and a History of the Thomsonian Materia Medica, As Shown in "The New Guide to Health" (1835), and the Literature of That Day (1909), by John Uri Lloyd and Samuel Thomson (PDF at swsbm.com)

Items below (if any) are from related and broader terms.

Filed under: Alternative medicine

Natural Liberty: Rediscovering Self-Induced Abortion Methods (Creative Commons licensed edition, c2008), by Sage-Femme Collective (Javascript-dependent Flash at scribd.com)

ID: how it's used Rameau

Concept information

URI	http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb14521343b
prefLabel	x-notation: FRBNF145213438 fr: Web sémantique
note	fr: Domaine : 621
inScheme	Rameau Rameau - Noms Communs
broader	Web
related	Ontologies (informatique) Services Web

Mappings (simple SKOS statements)

Mapping Relation	Concept
closeMatch	http://id.loc.gov/authorities/sh2002000569#concept

Additional vocabularies coming

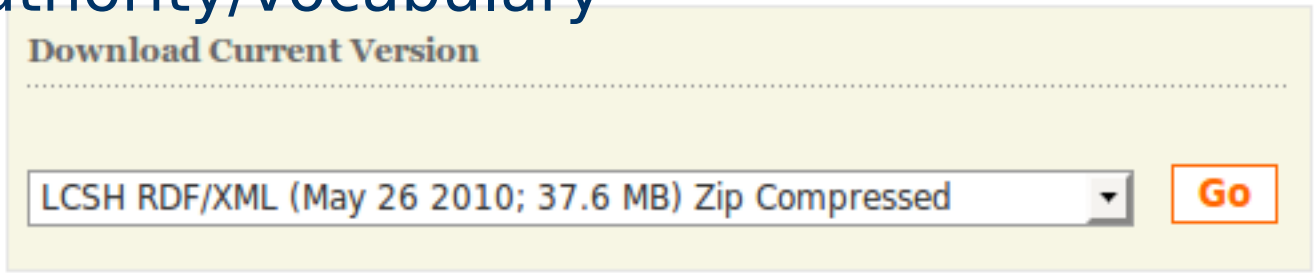
- ISO 639-2 and other ISO 639 standards (ISO 639-5 already in SKOS)
- MARC code lists (languages, geographic areas, countries)
- PREMIS controlled vocabularies
- Name authorities

Making LC's data available

- LC and Open Government directive
  - Executive branch order
  - LC is legislative branch
- Fostering openness
  - Adhere to standards
  - Make datasets available
  - Provide contact forms
  - LC has a long history of making data available

Getting the data

Bulk downloads of each authority/vocabulary



Individual resources via content negotiation

- XHTML/RDFa
- RDF/XML
- N-Triples
- JSON

LC's terms of service

Terms of Service

The Library of Congress has prepared this vocabulary terminology system and is making it available as a public domain data set. While it has attempted to minimize inaccuracies and defects in the data and services furnished, THE LIBRARY OF CONGRESS IS PROVIDING THIS DATA AND THESE SERVICES "AS IS" AND DISCLAIMS ANY AND ALL WARRANTIES, WHETHER EXPRESS OR IMPLIED.

Technical infrastructure

- Django (Python)
- LCSH
  - MySQL
  - SKOS RDF generated at time of request
  - Operates, more or less, as traditional relational DB
  - MARC mapped to relational DB tables
- Everything else
  - RDFlib (Python library, uses MySQL as triplestore)
  - Runs on triples
  - XML to SKOS RDF/XML before ingest
  - XSL, Xquery used

Next steps

- MADS OWL Schema to enable identification of facets within name and subject authorities: Aeronautics--Soviet Union—History
- Enhance existing vocabularies to show relationships between terms and vocabularies
  - Broader/narrower relator terms
  - Matches to other vocabulary terms (e.g. MARC vs. ISO 3166 country codes)

Questions?

Rebecca Guenther

rgue@loc.gov

Semantic Interoperability Workshop

July 22, 2010